Building a Gold Standard for a Russian Collocations Database

Maria Khokhlova
St. Petersburg State University
E-mail: m.khokhlova@spbu.ru

Abstract

In the last decade, linguists have become increasingly interested in corpus material, which allows for a fresh approach to the phenomena that have already been extensively described in academic works. The dual nature of the co-occurrence phenomenon itself lies, on one hand, in its linguistic component and, on the other, in the probabilistic (combinatorial) characteristics. The former has been described in numerous papers and explicitly defined in dictionaries, while the latter can be identified by a statistical approach. The present paper focuses on the process of building a gold standard that will include data from Russian dictionaries and corpora. The standard is being prepared for a Russian Collocations Database that already includes information on words' collocability and was extracted from text corpora by statistical measures and linguistic filters. The gold standard will be also used for the evaluation of the extracted collocations and for marking them as "true" collocations with references to the dictionaries.

Keywords: database, collocations, corpora, dictionaries, Russian language

1 Introduction

The study of a language assumes a large amount of data about word usage, and about their joint occurrence. The competence of a native speaker involves not only knowing numerous meanings of different words of the language, but also understanding which words, if connected, can form a single semantic unit. One can speak about a general population, which would cover absolutely all the examples available in a language. However, it is not possible to collect such a population for objective reasons. Nevertheless, information on collocability can be obtained from already created resources. Information on lexical and syntactic co-occurrence of various words can usually be found in dictionaries (explanatory, specialized, etc.), and, although less often, in grammar books. Large text corpora that have been actively appearing recently also can be seen as a source of such data. But they cannot serve *a priori* as a source of the relevant material without an appropriate "superstructure", as they contain typos, occasionalisms, errors and repetitions of the same fragments (such problems often arise when the data is automatically collected). As such, information on collocability that is automatically obtained on the basis of text corpora should be accompanied by some evaluation, allowing them it to be verified. It should be noted that there is no unified source of information on collocability that researchers can consult. Despite some criticisms, it should also be noted that dictionaries and other lexicographical resources are the most valuable sources of information on the collocability that must be used when preparing a gold standard. Therefore, the project is aimed at solving the following two tasks: 1) creating a single resource of "reference" (verified) data; 2) the representation of "reference" data on the basis of the text corpora, i.e., real language sets. Here we mean a kind of a "gold standard" of collocability, or a reference list containing as much information as possible about the word combinations and collocations for the Russian language. By a gold standard we understand a collection of collocations extracted from explanatory and specialized Russian dictionaries with some statistical evaluation measured on corpus data. The present paper describes an ongoing project and is structured as follows. The Introduction presents the basic idea of the research. Section 2 describes the related

projects and provides an overview of the topic. Section 3 discusses the main ideas underlying the gold standard and gives an example of its part. The last section concludes the paper and proposes plans for future work.

2 Related Work

Over the last decade, linguistics has been less focused on prescriptive tools, giving scholars more opportunities to draw their own conclusions based on the analysis of examples. Besides, the work of a lexicographer is believed to possess an inevitable element of subjectivity that influences the words and word combinations selected for the dictionary, as well as the way and order in which they are grouped into dictionary entries. It is not uncommon for a dictionary to become outdated in the time period between the start of its compilation and its publication date. It should also be mentioned that there is no single concept for data presentation in various explanatory dictionaries. For instance, the Russian noun "*nadezhda*" 'hope' has only the following standard collocations listed in the dictionaries (Kuznetsov 1998; Dictionary of the Russian Language 1981-1984): "*vozlagat nadezhdu*" 'to pin hopes', "*pitat nadezhdu*" 'to nourish hopes', "*podayot nadezhdu*" 'to give hope' and "*l'stit sebya nadezhdoy*" 'to flatter oneself with a hope'. These examples, however, do not include other collocations, which are also characteristic of this lexeme (for instance, "*opravdyvat nadezhdy*" 'to justify hopes' or "*vselyat nadezhdu*" 'to inspire a hope'). Therefore, the presence of such a non-uniform representation of collocational preferences of lexical units indicates that there is a need for a single resource that would absorb information from different sources.

There are a number of online systems developed for the Russian language, which provide information on collocations. Among these, we can name such resources as the Lexicograph database, the FrameBank database of collocations (Lyashevskaya 2010) which includes descriptions of valency frames for verbs and constructions, and the "Collocations, Colligations, Constructions" database (Kopotev et al. 2015), providing information about collocations on the basis of the Russian National Corpus (RNC) and the ruWac corpus. One can name dictionaries created on the basis of RNC as a source of lexicographic data. The RNC provides a range of instruments (n-gram search with statistical analysis, lists of established words and set expressions, lexical graphs), and has also served as a basis for compiling dictionaries. There is the *Dictionary of Verbal Collocations for Abstract Nouns of the Russian Language* (Biryuk, Gusev, & Kalinina 2008). It lists more than 1,000 collocations based on the following models: 1) noun + verb; 2) verb + noun; 3) verb + adjective + noun. We should also note the *Dictionary of Russian Idioms: Combinations of Words with the Meaning of High Degree* (Kustova 2008). Their differences from the proposed resource include the following: these dictionaries provide information on co-occurrence for a limited set of key words (for instance, only for verbs); they also offer an interface that is non-intuitive for a general user. Another unique lexicographic project is the *Active Dictionary of the Russian Language* compiled under the guidance of Yu. D. Apresyan, which offers vast amounts of information on co-occurrences separately represented in dictionary entries. The material is well-structured and includes data on syntactic actants, collocations and constructions. Another resource developed for the Russian language is the RNC Sketches project aimed at creating patterns of word sequences based on the material from the National Corpus of the Russian Language, which offers syntactic models for word sequences with examples but does not provide any quotations. Sketch Engine is another tool that provides information on syntactic co-occurrence on the basis of text corpora for different languages, including Russian (Kilgarriff et al. 2014; Khokhlova & Zakharov 2010).

Despite the fact that there are dictionaries of collocations, there exists, nevertheless, a need for resources that would describe the data more consistently. Currently there is no single system for the

Russian language which would allow scholars to obtain not only the corpus data, but also vocabulary information. At the moment there is no single system for the Russian language that would allow researchers to obtain not only corpus data, but also "reference" information on the vocabulary and behavior of lexical units. There are a number of unique and valuable lexicographic projects that describe the collocability, although in different ways. The purpose of our project is thus to consolidate the information on the collocability, which is presented in different ways in different sources. There are reference web-sites (for example, slovari.ru or gramota.ru), which provide an opportunity to learn the meaning of a word and to look through dictionary entries, which also contain information about collocability. But at the same time, users may have difficulty in reading the articles, as the information on collocability can be represented both in the "phraseological" part, and also directly in the quotations themselves. When it comes to collocations, a full dictionary entry with explanation is not always necessary, but rather examples of real data, appropriately designed and accompanied by values of the correctness or frequency of use in speech. Therefore, at the moment there is no such system, which would combine "reference" data from a lot of recognized sources, as well as real case examples. Such a system could be in demand both by ordinary users (for example, studying the Russian language) and by specialists. The given goal can be achieved right now, when large data sets are available (large text corpora), along with the software tools and computational power needed to process them.

3 Russian Collocations Database

3.1 Evaluation of Collocability

The task of creating corpora that comprise large amounts of data has a long history. Researchers have long been attracted by the opportunity to test their hypotheses on quantitatively new material, but only with the advent of new technologies has this been practical. Numerous works discuss different approaches to the automatic collection of material from the web (see, for example, Kilgarriff and Grefenstette (2003), Belikov et al. (2012), among others) and the creation of large text corpora. The threshold of 1 million tokens or even 100 million tokens has already been passed, and a number of papers discuss the merits of such large corpora (Belikov, Selegey, & Sharoff 2012; Benko & Zakharov 2016).

Along with the existence of active and passive vocabulary, the concept of active and passive dictionaries has also been introduced (Active Dictionary of the Russian Language 2014: 5-7). The first covers the needs of speaking and producing texts (ibidem). Thus, it gives a "deeper" idea of lexical units, describing not only their meanings, but also syntagmatic and paradigmatic properties. While the latter type of dictionary is aimed at covering as much material as possible, including low-frequency lexis. Large corpus of texts can be used to obtain data on collocability and its full representation within the paradigm of the active dictionary approach. The method used in the development of the Tolkovo-combinatorial dictionary (Melchuk & Zholkovsky 1984) showed the possibility of a structural approach to the description of collocability on the example of lexical functions.

When processing a large amount of data, it is difficult to use only a "manual" approach. Researchers thus work with automatic methods, which may include a quantitative approach, a rule-based approach, and a combination thereof. There are various statistical metrics for evaluating collocability. In other words, we are talking about the statistical non-randomness of word combinations, which can be evaluated quantitatively. Thus, some stability inherent to lexical units can be calculated, which allows them to be put on a scale: from free combinations to phraseological structures. Both statistical methods (including machine learning) and rules-based approaches can be used. In total, there are more than 80 measures to assess the strength of the relatedness of word combinations

(Pecina 2005). Not all of these were tested on language data, but the most frequently mentioned ones in the literature (MI, t-score, Dice, log-likelihood, chi-square) have proved successful when working on material from various different languages, including Russian. To evaluate the results of automatic collocation detection, both data from lexicographic sources are used (see, for example, Khokhlova (2008)) and the results of experiments with native speakers (Pivovarova et al. 2017). At the same time, there should be a lot of reference data to cover as many automatic results as possible, in order to evaluate them.

3.2 Representation of Information on Collocability

To build the gold standard, at the present stage we selected four explanatory Russian dictionaries (*Dictionary of Contemporary Literary Russian Language* 1948-1965; *Dictionary of the Russian Language* 1981–1984; *Big Academic Dictionary of Russian* 2004–2018; Kuznetsov 2014) and two specialized ones (Denisov & Morkovkin 1983; Borisova 1995). The given dictionaries differ in their representation of collocations and their example coverage. Explanatory dictionaries implement various ways to represent the information on combinatorial restrictions. One can find set phrases not only in special sections of the entrie,s but also in the examples, sayings and quotations. The entry structure can also vary and depends on a dictionary. For example, the diamond symbol ◊ is used to designate set expressions and phraseological units in the *Dictionary of the Russian Language 1981–1984*, while these are indicated in the *Big Academic Dictionary of Russian 2004–2018* with a tilde symbol.

We have collected collocations from the phraseological section in *Dictionary of the Russian Language 1981–1984*; the whole list comes about 13,000 examples (while the vocabulary list has more than 80,000 words). In the dictionary by Borisova (1995) collocations are structured according to their semantics and represented with a font. The analysis showed that both explanatory dictionaries (the *Dictionary of Contemporary Literary Russian Language 1948-1965* and *Big Academic Dictionary of Russian 2004–2018)* overlap with each other in their representation of collocations.

One single format implies that part-of-speech tags will be assigned to all the extracted collocations, as well as information as to in which dictionaries they were described.

In order to obtain data on co-occurrences in the Russian language, we process the Araneum Russicum Maximum corpus (with about 15 billion words), which was created automatically and is based on web texts of different genres, and is one of the largest collections of Russian texts (Benko & Zakharov 2016). We use a statistical approach for automatic extraction of word combinations from corpora that implied several association measures (t-score, MI, log-likelihood). We focused our attention on the bigrams within the span [-1; 1] from the node. Thus the following models were extracted: noun + verb, verb + noun, adjective + noun, noun + noun etc.

We analyzed phraseological sections of the entries marked with special symbols and extracted data from them. Then we merged two lists, hence there are three categories of collocations: 1) the overlapping collocations that have both references to dictionaries and statistical values (see Table 1); 2) collocations described in the dictionaries but not extracted from the corpus (as they can be longer than bigrams); 3) collocations that were not found in the dictionaries. In our study we focus on the first and second groups of word combinations.

Table 1 gives an example of collocations from the gold standard for the headword *"nadezhda"* ('hope') described in the dictionaries. The second column indicates the dictionaries that list a collocation. One can see that only one collocation is present in the entries of all the dictionaries ("*pitat nadezhdu*" 'to nourish hopes'), while other phrases were described in fewer lexicographic resources. As noted in Section 2, the coverage of specialized dictionaries can be even wider than that of

explanatory dictionaries. We introduced a simple metric called "dictionary index" that is given in the third column. It shows the number of dictionaries that include the collocation. It can vary within 0 and 6 (0 means that the collocation was not listed in any dictionary but nevertheless was extracted from the corpus). Large values of the index imply that the collocation is reproduced in speech quite often, and thus should be learned by heart (if we speak about students of Russian). The last three columns show the values of the association measures (t-score, MI, LL).

Table 1: Representation of the results for the headword *"nadezhda"* ('hope').

Collocation	Dictionaries[1]	Dictionary Index	Syntactic Structure	t-score	MI	LL
"*pitat' nadezhdu*" 'to nourish hopes'	1, 2, 3, 4, 5, 6	6	V+N	43,445	7,466	15991,086
"*podavat' nadezhdy*" 'to give hopes'	1, 2, 3, 4, 6	5	V+N	73,600	7,208	44053,620
"*podavat' nadezhdu*" 'to give a hope'	1, 3, 5, 6	4	V+N	73,600	7,208	44053, 620
"*pitat' nadezhdy*" 'to nourish hopes'	1, 2, 3, 4	4	V+N	43,445	7,466	15991,089
"nadezhda na" 'hope on'	1, 3, 5	3	N+Prep	405,451	3,796	682980,452
"v nadezhde" 'in hope'	1, 2, 3	3	Prep+N	229,645	1,842	118188,997
"vozlagat' nadezhdy" 'to pin hopes'	2, 4, 6	3	V+N	67,042	10,257	55472,800
"*vselyat' nadezhdu*" 'to inspire a hope'	5, 6	2	V+N	75,031	11,344	78560,080
"poslednyaya nadezhda" 'last hope'	3, 5	2	Adj+N	109,620	4,753	60613,064
"*vozlagat' nadezhdu*" 'to pin a hope'	3, 5	2	V+N	67,042	10,257	55472,800
"vyrazhat' nadezhdu" 'to express a hope'	5, 6	2	V+N	77,419	7,599	51871,075
"s nadezhdoy" 'with hope'	1, 3	2	Prep+N	148,848	2,044	51765,512
"ostavlyat' nadezhdu" 'to give up a hope'	5, 6	2	V+N	64,155	5,806	25996,563
"leleyat' nadezdhu" 'to cherish a hope'	5, 6	2	V+N	36,618	9,744	15561,769

It can be seen that the dictionaries list not only collocations but also constructions and colligations. At the present stage of the study we deal with lemmatized word combinations. This is a restriction if it comes to Russian, as certain phrases are used in a certain morphological form (e.g. *"tret'yego dnya"* 'the day before yesterday') and such preferences should be studied separately from other forms. Aspectual verb forms we considered as one item ("*podavat'"* 'to give' vs *"podat'"* 'to give'). The examples (see Table 1) suggest that the same headword can be used in both singular and plural forms in a collocation, but these phrases are not equally presented in the dictionaries (cf *"vozlagat' nadezhdy"* 'to pin hopes' and "*vozlagat' nadezhdu*" 'to pin a hope'). At the moment the collocations have the same statistical measures, but if we distinguish between word forms they will differ accordingly.

The above-mentioned second group of collocations has the following examples: *"obmanyvat' sebya*

1 Here we use the followings symbols: 1 (*Dictionary of Contemporary Literary Russian Language 1948-1965)*; 2 (*Dictionary of the Russian Language 1981–1984*); 3 (*Big Academic Dictionary of Russian 2004–2018*); 4 (Kuznetsov 2014); 5 (Denisov & Morkovkin 1983); 6 (Borisova 1995).

nadezhdoy" 'to disappoint oneself with a hope', *"teshit' sebya nadezhdoy"* 'to please oneself with a hope', etc. They do not have any statistical evaluation as the trigrams were not extracted from the corpus. But nevertheless the given collocations will be added to the gold standard.

The Russian Collocations Database is already partially available upon request on the web (http://collocations.spbu.ru), and includes information about lexical collocations with statistical evaluations. For our research we have developed a special database to store pairs of collocated words and their correlation values according to various collocation metrics. The database is implemented by means of the MySQL engine and consists of three main tables:

- words table;
- collocations table;
- metrics table.

The gold standard will help to distinguish between different types of collocations, e.g. high-frequency and specialized phrases.

4 Conclusion and Further Work

The paper describes work in progress. At the present stage the database includes automatically extracted collocations from a large web corpus. The collocations are marked according to their presence in the Russian dictionaries (gold standard).

Further work will be focused on extraction of collocations from quotations and other examples used in the entries, as they can contain significant data. Moreover, the analysis confirmed a need to find correlation between the ranking of collocations from the gold standard and their statistical coefficients.

The results of this research project can be used in courses on lexicology, morphology, and syntax of the Russian language; they will be helpful for compiling dictionaries and grammar books, as well as for teaching Russian. The proposed resource will also be useful for studying Russian as a foreign language. The obtained results can be used for machine learning in programs connected with automated language processing, for instance, in systems for automatic clustering of word combinations and disambiguation.

References

A Dictionary of the Russian Language [Slovar' russkogo jazyka v 4 tomakh]. (1981–1984). Yevgen'yeva, A. P. (ed.-in-chief). Vol. 1–4, 2nd edition, revised and supplemented. Moscow: Russkij jazyk.

Active Dictionary of the Russian Language [Aktivnyy slovar' russkogo yazyka]. (2014–2017). Apresyan, Ju. D. (ed.) Vol. 1-3. M.: Yazyki slavyanskoy kul'tury.

Belikov, V.I., Selegey, V.P., Sharoff, S.A. (2012) Preliminary considerations towards developing the General Internet Corpus of Russian [Prolegomeny k proyektu General'nogo internet-korpusa russkogo yazyka (GIKRYa)]. In *Computational linguistics and intellectual technologies*. Vol. 11 (18). Moscow: Izd-vo RGGU, pp. 37-49.

Benko, V., Zakharov, V. (2016). "Very large Russian corpora: New opportunities and new challenges." *Computational Linguistics and Intellectual Technologies* 15:22, pp. 79–93. Moscow: Izd-vo RGGU.

Big Academic Dictionary of Russian [Bolshoy akademicheskiy slovar v 30 tomakh]. (2004–2016). Moscow-Saint-Petersburg: Nauka.

Biriuk, O. L., Gusev, V. Iu., Kalinina, E. Iu. (2008). *Dictionary of Russian Abstract Nouns' Verbal Collocability. A Dictionary based on the Russian National Corpus* [Slovar' Glagol'noi Sochetaemosti Nepredmetnykh Imen Russkogo Iazyka. Slovar' na osnove Natsional'nogo Korpusa Russkogo Iazyka]. Accessed at: http://dict.ruslang.ru [15/05/2018].

Borisova, E. G. (1995). *A Word in a Text. A Dictionary of Russian Collocations with English-Russian Dictionary of Keywords* [Slovo v tekste. Slovar' kollokatsiy (ustoychivykh cochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov]. Moscow.

Denisov, P. N., Morkovkin, V. V. (comp. and ed.-in-chief), (1983). *An Academic Collocation Dictionary of Russian* [Uchebnyi slovar' sochetaemosti slov russkogo iazyka]. Moscow: Russkij jazyk.

Dictionary of Contemporary Literary Russian Language [Slovar sovremennogo russkogo literaturnogo yazyka v 17 tomakh], (1948–1965). Chernyshev, V.I. (ed.). Moscow-Leningrad: Izd-vo Akademii nauk SSSR.

FrameBank. Accessed at: http://framebank.ru/ [15/05/2018].

Khokhlova, M. (2008). Evaluation of Methods for Collocation Extraction [Eksperimental'naja proverka metodov vydelelnija kollokacij]. In *Slavica Helsingiensia 34. Instrumentarij rusistiki: Korpusnye podhody*. Eds. A. Mustajoki, M.V. Kopotev, L.A.Birjulin, J.J. Protasova. Helsinki. pp.343–357.

Khokhlova, M., Zakharov, V. (2010). Studying Word Sketches for Russian. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC'10) (Valetta, Malta, 19–21 May 2010). Eds. Nicoletta Calzolari (Conference Chair), Khalid Choukri, et al., pp. 3491–3494.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. In *Lexicography*, 1, pp. 7-36.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. In *Computational Linguistics*, 29 (3), pp. 333–347.

Kopotev, M., Escoter, L., Kormacheva, D., Pierce, M., Pivovarova, L., Yangarber, R. (2015). CoCoCo: Online Extraction of Russian Multiword Expressions. In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing* (10–11 September 2015, Hissar, Bulgaria). Sofia: INCOMA Ltd, pp. 43–45.

Kustova, G.I. (2008). Slovar' russkoj idiomatiki. Sochetaniya slov so znacheniyem vysokoy stepeni. [Dictionary of Russian Idioms: Combinations of Words with the Meaning of High Degree]. Accessed at: http://dict.ruslang.ru [15/05/2018].

Kuznetsov, S. (ed.). (2014). *Large Explanatory Dictionary of Russian* [Bolshoy tolkovyi slovar russkogo yazyka]. Norint, St. Petersburg.

Lexicograph. Accessed at: http://lexicograph.ruslang.ru/ [15/05/2018].

Lyashevskaya, O. (2010). Bank of Russian constructions and valencies. In *LREC* 2010. Malta, Valletta, May 19-21, 2010.

Mel'čuk, I., Zholkovsky, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian* [Tolkovo-kombinatornyj slovar russkogo jazyka]. Vienna.

Pecina, P. (2009). Lexical Association Measures. Collocation Extraction. Prague: Institute of Formal and Applied Linguistics.

Pivovarova, L., Kormacheva, D., Kopotev, M. (2017). Evaluation of collocation extraction methods for the Russian language. In Quantitative Approaches to the Russian Language (ed. by M. Kopotev, O. Lyashevskaya, A. Mustajoki). London, New York: Routledge. pp. 137–157.

Russian National Corpus. Accessed at: http://ruscorpora.ru [15/05/2018].

Acknowledgements

This work was supported by the grant of the President of Russian Federation for state support of scholarly research by young scholars (Project No. MK-2513.2018.6).